

(19)日本国特許庁 (JP)

## (12) 公表特許公報 (A)

(11)特許出願公表番号  
特表2001-500345  
(P2001-500345A)

(43)公表日 平成13年1月9日(2001.1.9)

(51)Int.Cl.<sup>7</sup>  
H 04 L 12/56  
12/28  
12/46

識別記号

F I  
H 04 L 11/20  
11/00マーク(参考)  
102D  
310C

審査請求 未請求 予備審査請求 未請求(全39頁)

(21)出願番号 特願平11-505717  
 (86) (22)出願日 平成10年6月24日(1998.6.24)  
 (85)翻訳文提出日 平成11年2月26日(1999.2.26)  
 (86)国際出願番号 PCT/US98/13205  
 (87)国際公開番号 WO99/00938  
 (87)国際公開日 平成11年1月7日(1999.1.7)  
 (31)優先権主張番号 08/885,114  
 (32)優先日 平成9年6月30日(1997.6.30)  
 (33)優先権主張国 米国(US)  
 (81)指定国 EP(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), JP

(71)出願人 サン・マイクロシステムズ・インコーポレーテッド  
アメリカ合衆国・94303・カリフォルニア州・パロ・アルト・サン・アントニオ・コード・901  
 (72)発明者 ヘンデル, エアリエル  
アメリカ合衆国・95014・カリフォルニア州・カッパチーノ・ニューキャッスル・ドライブ・7537  
 (74)代理人 弁理士 山川 政樹(外5名)

最終頁に続く

## (54)【発明の名称】多層分散ネットワーク要素におけるルーティング

## (57)【要約】

知られているルーティング・プロトコルに従ってパケットを中継する多層分散ネットワーク要素(201)。複数のサブシステム(201)の分散アーキテクチャはサブネットワーク間にまたがって回線速度のパフォーマンスでルーティングを実行する。それぞれのサブシステム(210)は転送用メモリ(213)および関連メモリ(214)を含み、ルーティングのためにユニキャストおよびマルチキャスト・パケットを識別し、VLAN情報の変更を含むハードウェア内のパケットの変更を行い、パケットを次のホップへ転送する。ルーティングの決定はインバウンド・サブシステム内で行われ、必要に応じてネットワーク・トポロジを付与されて個別のアウトバウンド・サブシステムを介してパケットが転送される。

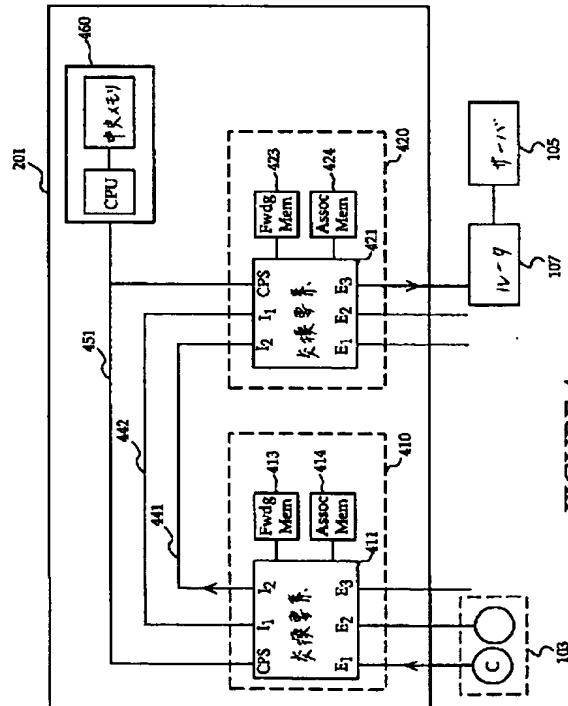


FIGURE 4

**【特許請求の範囲】**

1. 内部リンクで結合され、第1および第2の転送用メモリをそれぞれ備えた第1および第2のサブシステムを備えたノード間でパケットを受信して転送するネットワーク要素において、

第1のサブシステムによって第1のヘッダ部分を備えたパケットを受信するステップと、

前記パケットの第1のヘッダ部分の解析に応答してルーティング・プロトコルに従ってパケットを転送するかどうかを決定するステップと、

第1の転送用メモリに第1のパケットの第2のヘッダ部分に一致する第1のエントリがあるか検索するステップと、

前記パケットの第2のヘッダ部分の第1のエントリとの一致に応答して前記パケットの第1のヘッダ部分の一部を第1のエントリに関連付けられた隣接ノードの次のホップ・アドレスで置き換えるステップと、

次のホップ・アドレスを備えた前記パケットを内部リンクを介して第2のサブシステムへ送信するステップと、

前記パケットを隣接ノードへ転送するステップと  
を含むパケット中継の方法。

2. 第1のヘッダ部分が第2層の宛先アドレスを含み、第2のヘッダ部分が第3層の宛先アドレスを含む請求項1に記載の方法。

3. パケットを転送するかどうか決定するステップが第1のヘッダ部分がネットワーク要素に割り当てられたアドレスに一致するかどうかを決定するステップを含む請求項1に記載の方法。

4. 第2のヘッダ部分の第1のエントリとの一致に応答して内部リンクを介して第2のサブシステムへ制御信号を送信するステップと、

第2のサブシステムによって前記パケットを受信するステップと、  
第2のサブシステムの制御信号の受信に応答して前記パケットの第3のヘッダ部分を第2のサブシステムのアドレスで置き換えるステップ  
をさらに含む請求項1に記載の方法。

5. 第3のヘッダ部分が第2層の送信元アドレスを含む請求項4に記載の方法。
6. 前記パケットを第2のサブシステムへ送信するステップが、内部リンクを結合し第1のエントリに関連付けられた値として識別される第1のサブシステムの内部ポートへ前記パケットを送信するステップを含む請求項1に記載の方法。
7. パケットを転送するステップが第2のサブシステムのアドレスと共に前記パケットを転送するステップを含む請求項4に記載の方法。
8. 第1のパケットの第3のヘッダ部分を置き換えるステップが第3のヘッダ部分の一部を第2のサブシステムの第2の外部ポートのL2アドレスで置き換えるステップを含む請求項4に記載の方法。
9. 第1のサブシステムによってルーティング・プロトコルに従って前記パケットを転送することを決定するステップに応答して生存時間フィールドを更新するステップと、

第1のサブシステムによって前記パケットのヘッダ・チェックサムを補正するステップをさらに含む請求項1に記載の方法。

10. 前記パケットの第3のヘッダ位置の一部を置き換えた後で第2のサブシステムによって前記パケットの巡回冗長符号（C R C）を計算するステップをさらに含む請求項4に記載の方法。

11. 内部リンクを介したNEW TAG通知の受信に応答して第2のサブシステムの仮想ローカル・エリア・ネットワーク識別（V I D）を前記パケット内に挿入するステップをさらに含む請求項4に記載の方法。

12. ノード間でパケットを受信して転送するネットワーク要素であって、

第1の転送用メモリを備え、ネットワーク要素のアドレスに一致するパケットの第1のヘッダ部分の一部に基づいてパケットをルーティングするかどうかを決定するように構成された第1のサブシステムと、

第2の転送用メモリを備えた第2のサブシステムと、

第1および第2のサブシステムを結合し、

第1のサブシステムが前記パケットの第2のヘッダ部分の第1の転送用メモリ内の第1のエントリとの一致に応答して第1のヘッダ部分の一部を隣接ノードの次のホップ・アドレスで置き換えるように構成され、

第1のサブシステムが、さらに次のホップ・アドレスを備えた前記パケットを内部リンクを介して第2のサブシステムへ送信するように構成され、

第2のサブシステムが第1のヘッダ部分の第2の転送用メモリ内の第2のエントリとの一致に応答して受信した前記パケットを隣接ノードへ転送するよう構成される内部リンクと  
を有するネットワーク要素。

13. ネットワーク要素のアドレスが前記パケットを受信する外部ポートのL2アドレスである請求項12に記載のネットワーク要素。

14. ノード間でマルチキャスト・パケットを受信して転送するネットワーク要素であって、

第1の転送用メモリを備え、前記パケットのヘッダ部分に一致した転送用メモリ内の第1のエントリに基づいてマルチキャスト・パケットをルーティングするかどうかを決定するように構成された第1のサブシステムと、

第2のエントリを含む第2の転送用メモリを備えた第2のサブシステムと、

第1および第2のサブシステムを結合し、マルチキャスト・パケットを第1のサブシステムから第2のサブシステムへ転送するための内部リンクとを備え、

第2のサブシステムが内部リンクを介したマルチキャスト・パケットの受信と、第2のエントリの前記パケットのヘッダ部分との一致と、前記複数のパケットのそれぞれの第3のヘッダ部分の第2のサブシステムのアドレスでの置き換えとに応答して、複数のパケットを複数の隣接リンクへ転送するよう構成されたネットワーク要素。

15. 第1および第2のエントリがそれぞれネットワーク層のアドレスを含む請求項14に記載のネットワーク要素。

16. 前記複数のパケットのそれぞれの第3のヘッダ部分がデータ・リンク層の送信元アドレスを含む請求項14に記載のネットワーク要素。

17. 第2のサブシステムが第1のサブシステムから内部リンクを介して第2の制御信号を受信し、前記受信に応答して第2の転送用メモリの検索が実行される請求項14に記載のネットワーク要素。

18. 第2のエントリ内の検索結果が前記パケットのヘッダ部分に一致する請求

項17に記載のネットワーク要素。

19. 内部リンクが第1のサブシステムからの待ち行列化の優先度情報を第2のサブシステムへ転送するようにさらに構成された請求項14に記載のネットワーク要素。

20. 第3のエントリ内の検索結果がヘッダ部分に一致し、第2のサブシステムが第3のエントリに関連付けられた第3の制御標識をさらに含み、これに応答して第2のサブシステムが第1のサブシステムから受信した待ち行列化優先度標識を無効にする請求項17に記載のネットワーク要素。

## 【発明の詳細な説明】

### 多層分散ネットワーク要素におけるルーティング

#### 背景

##### 1. 発明の分野

本発明は一般的に言えば複数のコンピュータを結合する通信システム、より詳細にはネットワーク要素を介したメッセージ中継に関する。

##### 2. 関連技術の説明

コンピュータの相互通信は私的およびビジネス環境での日常生活の重要な態様になっている。コンピュータは、メッセージを送受信する物理媒体に基づいて、及びコンピュータに接続された電子ハードウェアとコンピュータで実行されるプログラムとによって実施される1組の規則に基づいて相互に通信を行う。しばしばプロトコルと呼ばれるこれらの規則は、接続された複数のコンピュータのネットワーク内でのメッセージの正常な送受信を決める。

ローカル・エリア・ネットワーク（LAN）は送信元コンピュータと宛先コンピュータの通信を可能にする最も基本的で最も簡素なネットワークである。LANは、相互に通信しようとするコンピュータ（端末局またはエンド・ノードとも呼ばれる）が接続されるネットワークと考えられる。少なくとも1つのネットワーク要素がLAN内のすべての端末局に接続される。簡素なネットワーク要素の一例はビットを転送する物理層のリレーであるリピータである。リピータはいくつかのポートを備え、それぞれのポートに各端末局が接続される。リピータは送信元端末局からのメッセージを含むデータ・パケットを形成しているビットを受信し、このパケットをビット単位でそのまま転送する。次に、これらのビットは宛先を含むLAN内の他のすべての端末局によって受信される。

しかしながら、单一のリピータへの物理的接続数は限られ、单一のリピータで処理できるメッセージの数にも限りがあるため、单一のLANでは多数の端末局

を備えた組織の要件を満たすには十分でない。したがって、これらの物理的制約が理由で、リピータベースの手法は限られた地理的領域で限られた数の端末局しかサポートすることができない。

しかしながら、コンピュータ・ネットワークの機能は、異なるサブネットワークを接続し、相互に通信する数千の端末局を含むより大きいネットワークを構成することで拡張されてきた。次にこれらのLANを相互接続して、ワイド・エリア・ネットワーク(WAN)リンクを含むさらに大きい企業ネットワークを構築することができる。

より大きいネットワーク内のサブネットワーク間の通信を容易にするため、より複雑な電子ハードウェアおよびソフトウェアが提案され、既存のネットワーク内で現在使用されている。また、適切なネットワーク要素で相互接続された端末局が、同じサブネットワーク内の端末局が共通の分類を備えるネットワーク階層を決めるという原則に基づいて、これらの端末局間の信頼性がある正常な通信のための新しい1組の規則が決められている。したがって、ネットワークはネットワーク内のノードおよび端末局の階層内の位置を定義するトポロジを備えると言われる。

パケット交換ネットワークを介した端末局の相互接続は、従来ピアツーピア階層アーキテクチャ概念に従っていた。このようなモデルでは、送信元コンピュータ内の所与の層がネットワーク内のピア端末局(通常は宛先)の同じ層と通信する。上位層から受信したデータ単位にヘッダを取り付けることで、層はその上位の層での動作を可能にするサービスを実施する。受信パケットは通常、送信元で動作する複数の異なる層によって元のペーロードに追加されたいくつかのヘッダを備える。

ARPANETおよびオープン・システム間相互接続(OSI)モデルなどのいくつかの層分割方式が従来技術には存在する。本発明を説明するためにここで用いる7層のODIモデルは他のモデルの機能性および詳細な実施態様を描くのに便利なモデルである。ただし、ARPANETのさまざまな態様(現在、Internet Engineering Task Force、すなわちIETFによって再定義されている)も下記の本発明の特定の実施態様に用いられる。

この場合の背景にある目的の該当する層は第1層(物理層)、第2層(データ・リンク層)、第3層(ネットワーク層)、および第4層(トランスポート層)

の一部である。これらの層に関連付けられた機能を以下に概説する。

物理層は通信リンク上で未構成の情報ビットを伝送する。リピータはこの層で動作するネットワーク要素の一例である。物理層自体はコネクタのサイズおよび形状、ビットの電気信号への変換、およびビット・レベルの同期などの問題を取り組む。

第2層はデータ・フレームの传送および誤り検出を規定する。より重要なことは、本発明で参照するデータ・リンク層は通常、单一ホップ、すなわちパケットのあるノードから他のノードへの移動距離を「ブリッジする」すなわち、单一ホップで情報パケットを搬送するように設計されている。最小時間を用いて受信パケットを処理してからパケットを次の宛先へ送信することで、データ・リンク層は次に述べる上位の各層よりもはるかに高速でパケットを転送することができる。データ・リンク層はデータ・リンク層またはその下位の層で相互接続されたあらゆるコンピュータ間の送信元および宛先を識別するためのアドレス指定を実現する。第2層のブリッジ・プロトコルの例はCSMA/CD、トーカン・バス、およびトーカン・リングなどのIEEE802で定義されたプロトコルを含む(Fiber Distributed Data Interface、すなわちFDDIを含む)。

第2層と同様に、第3層も相互に通信するコンピュータのアドレスを付与する機能を備える。しかしながら、ネットワーク層はネットワーク階層に関するトポロジ情報も扱う。またネットワーク層は最短経路を用いて送信元から宛先へパケットを「ルーティング」するように構成できる。最後に、ネットワーク層は、送信元がパケット速度を低減する要求として認識する選択したパケットをドロップする処理だけで輻輳を制御できる。

最後に、第4層のトランスポート層はアプリケーションが第3層とインターフェースをとるのに用いる「ポート・アドレス」を備えた電子メール・プログラムなどのアプリケーション・プログラムを実行する。トランスポート層とその下位の層の主要な差は送信元コンピュータ上のプログラムが宛先コンピュータ・システム上の同様のプログラムとの通信内容を伝送する一方で、下位の層では、プロト

コルはあるコンピュータとそれに隣接したコンピュータ間に用いられ、最終的な

送信元および宛先端末局はいくつかの中間ノードによって分離される場合があるということである。第4層および第3層のプロトコルの例はTCP（伝送制御プロトコル）およびIP（インターネット・プロトコル）などのプロトコルのインターネット・スイッチ（switch）を含む。

端末局はパケットの送信元および最終宛先であり、ノードは端末局間の中間点をさす。ノードは通常パケット単位でメッセージを受信して転送する機能を備えたネットワーク要素を含む。

一般的に言えば、規模が増大し複雑になったネットワークは通常上位層（第3層および第4層）の機能を備えたノードに依存する。いくつかのより小さいサブネットワークからなる非常に大きいネットワークは通常サブネットワークのトポロジを認識できるルータとして知られた第3層のネットワーク要素を用いる必要がある。

ルータはその隣接ノードとの情報交換に基づいてその周囲のネットワークのトポロジ・マップを形成して記憶することができる。LANの設計上、第3層のアドレス指定機能が含まれる場合、ルータを用いて端末局から得られる階層ルーティング情報を利用することで複数のLAN間でパケットを転送できる。端末局のアドレスおよび経路のテーブルがルータによってコンパイルされると、ルータが受信したパケットは、そのパケットの第3層の宛先アドレスをメモリ内の既存の一致するエントリと比較した後で転送できる。

ルータは、受信パケットのヘッダを解析し、ルータ内部のルーティング・テーブルに基づいて決定を下し、次のノードまたは端末局へ必要なヘッダの修正を加えてパケットを転送するように動作する。したがって、パケットは宛先に到着する前にこのようないくつかの「ホップ」を経る。ホップは、1つのノードまたは端末局から別のノードまたは端末局へ移動するパケットとして定義される。

ルータと比較して、ブリッジは第3層ではなくデータ・リンク層（第2層）で動作するネットワーク要素である。ブリッジは通常、メディア・アクセス制御（MAC）アドレスと呼ばれるパケットの宛先の第2層のアドレスにのみ基づいてパケットを転送する。一般的に言えば、ブリッジはパケットに変更を加えない。

ブリッジは端末局からの協力がない、階層を備えていない平坦なネットワーク内でパケットを転送する。

ブルータおよび交換などのハイブリッド形態のネットワーク要素も存在する。ブルータはブリッジの機能も兼ね備えたルータである。交換という用語はそれぞれの機能が汎用のプロセッサが実行する命令ではなくハードワイヤード論理で実施される高速のパケット転送が可能なネットワーク要素をさす。さまざまな形態の交換は、第2層および第3層で動作する。

以上、現在のネットワーク技術一般について説明してきたが、以下にこのような従来技術の限界について説明する。今日のインターネット上で実行されるマルチメディア・アプリケーションのために既存のネットワークで利用できる帯域幅を増加させる必要があるユーザがますます増えるにつれて、現代の、および将来的のネットワークは極めて広い帯域幅と多数のユーザをサポートすることができなくてはならない。さらに、このようなネットワークは通常異なるサービス特性を必要とする音声およびビデオなどの複数のトラフィック・タイプをサポートできなくてはならない。統計によると、ネットワーク・ドメイン、すなわち相互接続されたLANのグループとそれぞれのLANに接続された個々の端末局の数は将来、これまでより速い速度で増加する。したがって、帯域幅の拡大とリソースのより効果的な利用がこれらの要件を満たすために必要である。

ブリッジなどの第2層の要素を用いたネットワーク構築によって、複数のLAN間のパケット転送は高速化されるが、トラフィック分離、冗長トポロジ、待ち行列化およびアクセス制御の終端間ポリシーの柔軟性は失われる。

サブネットワークの端末局は第2層または第3層のアドレス指定に基づいて会話を起動できる。ブリッジが第2層解析にのみ基づいてパケットを転送するため、端末局は簡素であるが高速の転送サービスを実現する。しかしながら、ブリッジは同じサブネットワーク内の端末局間の待ち行列化、優先度、および転送の制約を含む上位層処理ディレクティブの使用をサポートしていない。

サブネットワーク内のブリッジを用いた会話を拡張する従来技術のソリューションは第2層および上位層のヘッダの組み合わせを用いるネットワーク要素に依存する。前記システムでは、固定サービス品質（QoS）を備えた転送用メモリ

内の新しい第2層のエントリを用いて初期パケットの第3層および第4層の情報が検証され、パケットの「フロー」が予測されて識別される。その後、第2層のヘッダと転送用メモリ内の第2層のエントリとの一致に基づいて後続パケットが第2層の速度で（固定QOSを伴って）転送される。したがって、フローの識別のために第3層および第4層のヘッダを備えたエントリが転送用メモリ内に入れられることはない。

しかしながら、電子メール・プログラムやビデオ会議セッションなどの同じ1対の端末局間で送受信される複数のプログラムがあるシナリオを考えてみる。これらのプログラムが異なるQOSのニーズを抱えている場合、上記の従来技術の方式は転送時に第3層および第4層の情報を考慮しないため、同じ1対の端末局間での異なるQOS特性をサポートしない。この結果、同じサブネットワークに接続された端末局で実行されているアプリケーションからの個々の優先要求をサポートできる柔軟性を備えたネットワーク要素が必要になる。

後者の属性はルータなどの第3層の要素を用いて満足できる。ただし、ルータによってインテリジェンスおよび意志決定機能が向上した代わりにパケット転送速度が犠牲にされる。したがって、第2層および第3層の要素の組み合わせを用いてネットワークが構築される場合が多い。

インターネットを用いるブラウザ・ベースのアプリケーションに伴ってサーバの役割は増し、その結果トラフィック分散は多様化した。サーバの役割が例えばファイル・サーバに狭く限定されていた時には、クライアントとファイル・サーバを同じサブネットワークに配してネットワークを設計し、ルータのボトルネックを回避していた。しかしながら、ワールド・ワイド・ウェブおよびビデオ・サーバなどのより専門化したサーバは通常クライアントのサブネットワークにはなく、この場合、複数のルータを経由することが避けられない。したがって、パケットが複数のルータを高速で横断する必要が無視できない。ブリッジかルータかの選択は通常かなりのトレードオフとなる、すなわち、ブリッジを選べば機能が低下し、ルータを選べば速度が低下する。さらに、トラフィック・パターンがルータを含む場合はサーバの性能がサーバの位置で変わるために、あるネットワーク内のサーバ特性はもはや同種のものではない。

したがって、トポロジおよびメッセージ・トラフィックなどの変化するネットワーク条件に対処しながらパケットの第2層、第3層、および第4層のヘッダに基づいてパケットを交換する高性能ハードウェアを効率的に用いることができるネットワーク要素が必要である。このネットワーク要素はブリッジ同様の速度で動作しながら異なるサブネットワーク間でパケットをルーティングして品質サービスなどの上位層の機能を実現する必要がある。

### 概要

本発明は知られているルーティング・プロトコルに従って多層分散ネットワーク要素によるパケットの中継装置およびそれに関連する方法である。

本発明は知られているルーティング・プロトコルを用いてパケットを受信して転送する多層分散ネットワーク要素（MLDNE）を対象とする。MLDNEは内部リンクで結合されたいくつのサブシステムを備える。それぞれのサブシステムは転送用メモリと関連メモリを備える。これらのメモリはアドレスを含むパケット・ヘッダ情報をルーティング情報に関連付ける。サブシステムはまた隣接ノードおよび端末局に接続する外部ポートと、内部リンクを介して他のサブシステムに接続する内部ポートを含む。

パケットが第1の「インバウンド」サブシステムによって受信されると、このサブシステムは、MLDNEの第2層のアドレスに一致する受信パケットの第2層の宛先アドレスを含む第1のヘッダ部分に基づいてパケットをルーティングするかどうか決定する。受信パケットの第1のヘッダ部分がMLDNEアドレスに一致する場合、第1のサブシステムはその転送用メモリを用いて、受信パケットの第3層の送信元および宛先アドレスを含む第2のヘッダ部分について経路がすでに決定されているかどうか判定する。

転送用メモリ内のタイプ2エントリが受信パケットの第2のヘッダ部分に一致する場合、隣接ノードの第2層のアドレス（関連メモリ内にある）がパケットの第2層の宛先アドレスに取って代わる。隣接ノードのアドレスは、一致するタイプ2のエントリに関連付けられたルーティング情報の一部として関連メモリ内に以前に記憶されている。サービス品質情報に加えて、関連メモリ内のルーティン

グ情報も隣接ノードに接続するインバウンド・サブシステムの外部ポートを識別する。隣接ノードがインバウンド・サブシステム以外のサブシステムに接続されている場合、この状況は一致するタイプ2のエントリが作成された時点で認識されていたはずで、関連メモリが隣接ノードまたは端末局が接続されている他のサブシステムに接続する、外部ポートではなく、インバウンド・サブシステムの内部ポートを識別するであろう。

パケットが内部リンクを介して第2のサブシステムによって受信されると、このパケットは第2の転送用メモリ内のタイプ1のエントリに一致するパケットの新しい第1ヘッダ位置に応答して隣接ノードへ転送される。第2のサブシステム内のタイプ1のエントリは隣接ノードまたは端末局のアドレスを含み、インバウンド・サブシステムの一致するタイプ2のエントリとは無関係に作成されていた。

受信パケットをルーティングする決定をした後で、インバウンド・サブシステムはまた最終的にパケットを転送する外部ポートに対して隣接ノードへパケットを送信する前にパケットの送信元を識別する第3のヘッダ位置を変更するよう指示する第1の制御信号を生成する。パケットの第2層の送信元アドレスは外部ポートに関連付けられた送信元アドレスと置き換えられる。隣接ノードがそのサブシステムを介して到達可能である場合、制御信号はまた内部リンクを介して第2のサブシステムへ送信される。

本発明の分散アーキテクチャはまたマルチキャスト・パケットのルーティングをサポートするように構成できる。マルチキャスト・ルーティング可能なパケットがインバウンド・サブシステムによって識別されていると、第2の制御信号を内部リンクを介して送信でき、その応答として第2のサブシステムが転送用メモリ内のタイプ2検索を行う（パケットのネットワーク層および上位ネットワーク層のヘッダに基づいて）。一致するタイプ2のエントリが見付かった場合、第2のサブシステムの外部ポートが第1の制御信号（やはりインバウンド・サブシステムから受信した）をチェックしてパケットの送信元アドレスを置き換える必要があるかを確認し、次にヘッダを適当に変更してパケットが転送される。第1の制御信号はマルチキャスト宛先グループがインバウンド・サブシステムに接続さ

れたノード／端末局を備える前記インバウンド・サブシステムの外部ポートによ

って受信されチェックされることもある。

本発明の検索エンジン、転送エンジン、およびデータ構造は受信パケットについてルーティング判定基準が満たされない場合、ブリッジング機能が自動的に起動されるブリッジングおよびルーティング機能を同時にサポートする方法で組織される。

本実施形態で、本発明はデータ・リンク層（第2層）、ネットワーク層（第3層）、およびトランスポート層（第4層）を含む上位層で実施される。

#### 図面

本発明の前述の態様およびその他の特徴は以下の図面、詳細説明、および請求の範囲から明らかになろう。

第1図は本発明の多層分散ネットワーク要素（MLDNE）のネットワーク適用例の高レベル図である。

第2図は本発明の一実施形態として（MLDNE）の内部の図である。

第3図は本発明の別の実施形態によるパケットのルーティングのための関連付けられたデータを含むMLDNE内のサブシステムの転送用および関連メモリの一例を示す。

第4図は2つしかサブシステムを備えずクライアントとサーバ間のルータとして動作するMLDNEの一実施形態のブロック図である。

第5図は本発明のネットワーク要素によるルーティングのための受信パケットの処理の流れ図である。

第6図は第5図の流れ図の続きで、ユニキャスト・パケットの処理の際に実行されるステップを含む。

第7図はユニキャスト・パケットをルーティングする本発明のネットワーク要素によって実行されるステップおよび動作の例を示す。

第8A図は本発明の一実施形態で使用されるパケット構造の簡素なブロック図である。

第8B図は本発明によるパケットのヘッダ・フィールド置き換えのための構造

である。

#### 詳細な説明

例示の図面に示したように、本発明はいくつかのノードと端末局をさまざまな異なる方法で相互接続するネットワーク要素を定義する。特に、多層分散ネットワーク要素（MLDNE）の適用例では IEEE 802.3 標準またはイーサネットなどの同種データ・リンク層で事前定義されたルーティング・プロトコルに従ってパケットをルーティングする。第1図に MLDNE 201 がクライアント C をルータ 107 に結合し、次にルータ 107 がサーバ 105 に結合するネットワーク内での本発明のルータとしての使用を示す。MLDNE 201 はいくつかのデスクトップ・ユニット（端末局）を相互接続する一方でその外部接続 217 を介して中間ノードとして動作できる。MLDNE 201 はルータとして動作する一方で、サーバ 105 とクライアント C が異なる LAN に常駐する高性能通信経路をサーバとデスクトップ間に設定できる。

MLDNE の分散アーキテクチャは RIP および OSPF のようないくつかの知られているルーティング・アルゴリズムに従ってメッセージ・トラフィックをルーティングするように構成できる。好ましい実施形態では、MLDNE はプロトコルのインターネット・スイッチ、より詳細に言えば、伝送制御プロトコル（TCP）およびインターネット・プロトコル（IP）をイーサネットの LAN 標準およびメディア・アクセス制御（MAC）データ・リンク層で用いてメッセージ・トラフィックを処理するように構成される。この場合、TCP はまた第 4 層のプロトコルの例として参照され、IP は第 3 層のプロトコルの例として繰り返し参照される。ただし、本発明の概念を実施するためにその他のプロトコルを用いることもできる。

本発明の MLDNE の第 1 の実施形態では、ネットワーク要素がパケット・ルーティング機能を分散方式で実施する、すなわち、ある機能の異なる部分が MLDNE 内の同一のビルディング・ブロック・サブシステムによって実行される一方で全機能の最終結果が外部ノードおよび端末局からはトランスペアレントのままであるように構成される。以下の説明と第 2 図から理解できるように、MLD

N E は設計者がサブシステムをさらに追加することで外部接続の数を増すことを

可能にするスケーラブルなアーキテクチャを備える。

第2図にブロック図の形式で示すように、MLDNE 201 はより大きいネットワーク要素を作成するためのいくつかの内部リンク 241 を用いて完全にメッシュ化され相互接続されたいいくつかの同一のサブシステム 210 を備える。少なくとも1つの内部リンクが2つのサブシステムを結合する。それぞれのサブシステム 210 は転送用メモリ 213 と関連メモリ 214 を含む。転送用メモリ 213 は受信パケットのヘッダとの一致に用いるアドレス・テーブルを記憶する。関連メモリ 214 はパケットを MLDNE 内で転送するための転送属性を識別するための転送用メモリ内の各エントリに関連付けられたデータを記憶する。入出力機能を備えたいいくつかの外部ポート（図示せず）が外部接続 217 とのインターフェースになる。それぞれのサブシステム内の同様に入出力機能を備えた内部ポート（図示せず）は内部リンク 241 を結合する。好ましい実施形態では、外部および内部ポートは特定用途向け IC (ASIC) によって実施されるハードウェア論理交換要素 211 内に配置される。

受信パケットは外部接続 217 の1つを介してインバウンド・サブシステムに到着し、アウトバウンド・サブシステム内の別の外部接続を介して MLDNE 外のノードまたは端末局へ転送される。アウトバウンドおよびインバウンド・サブシステムは同じサブシステムの場合もあるし、異なるサブシステムの場合もある。

第2図を参照すると、MLDNE 201 は周辺コンポーネント相互接続 (PCI) などの通信バス 251 を介して個々のサブシステム 210 に結合される中央処理システム (CPS) 260 を含む。CPS 260 は中央メモリ 263 に結合された中央処理装置 (CPU) 261 を含む。中央メモリ 263 はさまざまなサブシステムの個々の転送用メモリ 213 に含まれるエントリのコピーを含む。CPS はそれぞれのサブシステム 210 への直接の制御および通信インターフェースを備える。また CPS は受信パケットを通常パケットの第3層の宛先アドレスに規定される最終的な宛先へ転送する経路の一部として隣接ノードを識別するため

のいくつかのルーティング・プロトコルで構成される。CPS260のその他の分担機能は異なるサブシステム間にパケット・バッファなどのデータ経路リソースを設定することを含む。最後に、CPS260はタイプ2のエントリをそれぞれ

の個々のサブシステムの転送用メモリに追加するかどうかを決定するという重要なタスクを実行する。

第3図はそれぞれのサブシステムの転送用メモリと関連メモリの詳細図である。転送用メモリはタイプ2のエントリ321とタイプ1のエントリ301の2つのタイプのメモリをいくつか含む。転送用メモリ内のそれぞれのエントリは受信パケットのヘッダと比較されるデータを含む。TCP/IPの特定の実施形態では、それぞれのタイプ2のエントリ321のデータ・フィールドはクラス・フィールド323、IP送信元フィールド325、IP宛先フィールド327、アプリケーション送信元ポート333、アプリケーション宛先ポート335、およびインバウンド・ポート・フィールド337を含む。タイプ1のエントリ301に関して、クラス・フィールド、第2層のアドレス・フィールド、およびVLAN識別フィールド(VID)が例示の実施形態に示されている。代替のネットワークおよびトランsport層のプロトコルを用いた追加のヘッダ情報および同様の定義を開発してそれぞれのエントリに含め、受信パケットのヘッダの突き合わせに使用できることは当然であり、当業者には明らかであろう。

それぞれのタイプ2のエントリ321とタイプ1のエントリ301に関連メモリ214に記憶された関連付けられたデータが関連する。関連データ・フィールドはサブシステムが受信した一致パケットを転送するために必要な情報を含む。サブシステム・ポート・フィールド347は一致パケットを次のホップで隣接ノードへ転送するために用いるサブシステムの内部または外部ポートを識別する。次のホップ・アドレス・フィールド357はルーティングされる受信ユニキャスト・パケットの元の第2層の宛先アドレスに取って代わる隣接ノードの第2層のアドレスを識別する。優先フィールド345は実際にパケットをMLDNE外に送信する外部ポートによって待ち行列化のために用いられる。エージ・フィールド343および344は最近受信したパケットが対応するタイプ1またはタイプ

2のエントリに一致したことと示すことで転送用メモリ内のエントリ数を最小化するのに役立つ。

NEW VIDアドレス・フィールド353を用いて仮想LAN(VLAN)をサポートするようにMLDNEを構成できる。関連付けられたデータは、サブ

システムに特にサブネットワーク間にまたがってパケットを転送する場合にパケットのVIDを変更する必要を知らせるNEW VLAN識別(VID)TAGフィールドも含む。これに応答してインバウンド・サブシステムは新しいタグを挿入するか、既存のタグをNEW VIDフィールドの値で置き換える。例えば、VLAN間のルーティングで転送されたパケットのタグが受信パケットのタグと異なることが必要な場合、NEW VIDフィールドはサブシステムがパケットの転送前に置き換える交換用タグを含む。

パケットが内部リンク上で送信される場合、パケットを受信するアウトバウンド・サブシステムは内部リンク上で追加の制御情報を利用できる。以下に説明するsa\_replaceに加えてこのような情報は受信パケットに元々VLAN情報のタグが付いていたかどうかを示すorig\_tagビット、タグがインバウンド・サブシステムによって変更されたかどうかを示すmod\_tagビット、およびアウトバウンド・サブシステムが受信パケットにタグを付けてはならないことを示すdon't\_tagビットを含む。

最後に、関連メモリは以下に詳述するサブシステム内のマルチキャスト・ルーティング機能を起動するマルチキャスト経路フィールド355を含むように構成できる。

MLDNE 201のルーティング動作を第4図のネットワーク適用例と関連して第5図～第7図の流れ図を用いて実施形態の一例について説明する。転送用メモリおよび関連メモリのフィールドへの参照は第3図にある。以下の例では、パケットの移動はMLDNE 201の外部接続に結合されたサブネットワーク103内のクライアントCから追跡される。クライアントCはパケットのヘッダの第3層の宛先アドレスで識別されるサーバ105へパケットを送信する。パケットはMLDNE 201によって知られる第2層のアドレスを備えていると思われる

ルータ 107 を通過する必要がある。

第5図のブロック 503 から始めて、パケットはインバウンド・サブシステム 410 の外部ポート  $E_1$  で MLDNE 201 によって受信される。パケットはローカルに定義されたネットワークのサブネットワーク 103 内に第3層のアドレスを備えたクライアント C から生成されたメッセージを含む。サブシステム 41

0 は外部ポート  $E_1$  および  $E_2$  がサブネットワーク 103 を結合していることを認識するように構成できる。

パケットが交換要素 411 によって受信されると、動作は判定ブロック 507 へ移り、本発明の第2層の宛先アドレスを含む受信パケットの第1のヘッダ部分が MLDNE 201 のルータ・アドレスと比較される。ルータ・アドレスは、外部ポート  $E_1$  に割り当てられた第2層アドレスでも、MLDNE 全体に割り当てられた第2層アドレスでもよい。一般に、MLDNE はそれぞれの外部ポートに専用のルータ・アドレスが割り当てられるように構成される。

受信パケットの最初のヘッダ部分がルータ・アドレスと一致すると、動作はブロック 515 へ進み、パケットは潜在的なユニキャスト経路の候補であることが宣言される。しかしながら、第1のヘッダ位置がルータ・アドレスに一致しない場合、動作はブロック 509 へ進み、パケットはユニキャスト・ルーティング可能パケットではないと宣言される。下記から分かるように、このようなパケットはやはり MLDNE 内で利用可能なマルチキャスト経路を備えたマルチキャスト・パケットの場合もある。

経路クラスのユニキャスト・パケットの場合、ブロック 517 はクラス・フィールド 323 に「route」を用いて転送用メモリ 413 内に一致するタイプ 2 のエントリがあるかを検索する。

ブロック 517 での転送用メモリの検索の結果、判定ブロック 521 でタイプ 2 の一致するエントリが転送用メモリ 413 内に存在するかが判定される。存在しない場合、動作はブロック 523 へ進み、受信パケットのヘッダの該当部分がサブシステム 410 の CPS ポートおよび CPS バス 451 を介して CPS へ送信される。

C P S 4 6 0 がブロック 5 3 3 でサブシステム 4 1 0 から「不明」パケットのヘッダの部分を受信すると、C P S は C P S と C P S 第 2 層および第 3 層のトポロジ・テーブル内に事前構成されているアクセス・ポリシーおよびサービス・ポリシーを検証する。C P S は受信パケットが要求した経路へのサービスを拒否して専用のソフトウェア内のみでルーティングを実行するか、その経路のインバウンド・システムの転送用メモリ内にタイプ 2 のエントリを作成する任意選択が可能である。

MLDNE201 のルーティング・アルゴリズムは C P S によって実施される。受信パケットについてユニキャスト経路が存在するか直ちに計算できる場合、C P S は判定ブロック 5 3 7 でブロック 5 3 9 へ進み、インバウンド・サブシステム 4 1 0 の転送用メモリに経路クラスのタイプ 2 のエントリ 3 2 1 を追加し、前記サブシステムの関連メモリに関連付けられたデータを追加する。中央メモリの第 2 層のテーブルに問い合わせる C P S が判定するように、隣接ノードがインバウンド・サブシステム 4 1 0 の外部ポートに接続する場合、外部ポートは新しいタイプ 2 のエントリの関連付けられたサブシステム・ポート・フィールド 3 4 7 内で外部ポートが識別される。同様に、隣接ノードがサブシステム 4 2 0 に接続する場合、内部ポート I<sub>1</sub> または I<sub>2</sub> が識別される。

判定ブロック 5 2 1 に戻って、パケットがインバウンド・サブシステム 4 1 0 の転送用メモリ 4 1 3 内の既存の経路クラス・タイプ 2 のエントリに一致する場合、受信パケットは第 6 図に例示形式で示すユニキャスト・パケットとして転送される。

第 6 図を参照すると、インバウンド・サブシステム内では、交換要素 4 1 1 がユニキャスト・パケットの生存時間を超過したかどうかを評価する。生存時間フィールドは受信パケットのヘッダ内に存在するものとする。生存時間フィールドが示すようにパケットがネットワーク内をあまりに長時間にわたって循環していると、インバウンド・サブシステムだけが受信パケットを C P S へ送信し、次に例えば内部制御メッセージ・プロトコル (ICMP) に準拠した、またはインターネット・コミュニティが維持するコメント要求 (RFC) に記述される時間超

過エラー・メッセージがブロック609でCPSによって生成される。

他方、パケットの生存時間 (TTL) を超過していない場合、動作はブロック619へ進み、TTLがデクリメントされる。パケットのヘッダのこの変更は一般にブロック621でのパケットの第3層のヘッダ・チェック・サムの補正を必要とする。ブロック611で、交換要素411は受信パケットの第2層の宛先アドレスを第5図のブロック521で判定された一致するタイプ2のエントリに対応する関連メモリにある次のホップの第2層のアドレスに置き換える。

MLDNE201がVLANをサポートするように構成されている場合、判定ブロック615はNEW VIDタグ・フィールド351のステータスをチェックすることで新しいVLAN識別タグが必要かどうかを判定する。

別のサブシステムによってパケットをMLDNEの外に転送してもしなくても（一致するタイプ2のエントリに関連付けられたサブシステム・ポート・フィールド347によって示されるように）、sa\_replaceビットなどの第1の制御信号がブロック621で生成される。sa\_replaceビットはサブシステム・ポート・フィールド347で示される外部および内部ポートへハンドオフされ、この結果、パケットと共に内部リンク441を介してサブシステム420へ転送できる。第1の制御信号はサブシステム（インバウンドのサブシステムまたは別のサブシステム）に第2層の送信元アドレスをパケットの転送に用いる外部ポートの送信元アドレスで置き換えるよう通知する。

第4図の例では、ブロック627で、第1の制御信号を含むあらゆる制御情報と共にパケットは交換要素411内の内部ポートI<sub>2</sub>によって処理され、アウトバウンド・サブシステム420に接続する内部リンク441へ送達される。ただし、別法として、変更パケットおよび制御情報はインバウンド・サブシステム内にとどまり、外部ポートによって処理され、動作はブロック630へ進む。

ブロック627で、パケットはアウトバウンド・サブシステム420内の内部リンクを介して受信される。タイプ1の突き合わせサイクルが開始し、判定ブロック629に達して転送用メモリ423内に一致するタイプ1のエントリが存在するかどうか判定される。タイプ1のエントリが存在する場合、動作はブロック

630へ進む。

ブロック630からブロック637までの動作は「アウトバウンド」サブシステムによって実行され、インバウンド・サブシステム410であろうが異なるサブシステム420であろうがパケットはMLDNEを離れる。判定ブロック630でチェックしたsa\_replaceビットが設定されると、交換要素が受信パケットの少なくとも第2層の送信元アドレスを含む第3のヘッダ位置をパケットを転送するための外部ポートE<sub>3</sub>の第2層のアドレスで置き換える。外部ポートE<sub>3</sub>はブロック629で見つかった一致するタイプ1のエントリ（パケットは

内部リンクを横断して到着した）またはブロック521で見つかった一致するタイプ2のエントリ（パケットはインバウンド・サブシステムに残っていた）に対応する（関連メモリ内の）関連付けられたデータ内で識別された。

MLDNEはそれぞれの外部ポートが一意的な第2層のアドレスを割り当てられるように構成できる。別法として、単一の送信元アドレスをMLDNE全体に割り当てることができる。いずれの場合も、第3のヘッダ部分の置き換えに続けてパケットのヘッダの巡回冗長符号（CRC）がブロック635で再計算され、パケットは第4図のルータ107である隣接ノードへ転送される。

上記の例で、MLDNE201のクライアントCに始まりサブシステム410、内部リンク441、およびサブシステム420に至るパケットの移動について説明してきた。次にパケットはルータ107によって受信され、従来の手段に従ってサーバ105へ転送される。上記はもちろんルータ107を介した宛先としてのサーバ105への経路が経路を決定する従来の技法を用いてMLDNE201によってすでに捕捉されていることを前提としていた。

上記はまた、ユニキャスト・パケットが経路クラス内にあるがパケットをMLDNEを介してルーティングするためのタイプ2の一一致するエントリがインバウンド・サブシステム内に存在しない状況もカバーしていた。したがって、受信パケットがルーティングされるかどうかの決定はインバウンド・サブシステム内、特に第5図の判定ブロック507および521で実行される。ルーティング・ポリシーおよびサービス・クラスの待ち行列化が第3層の終端間アドレスおよびプ

プロトコル・ベースの分類の細分性および柔軟性を備えることに注意されたい。これらのルーティング・ポリシーおよびサービス・クラスの待ち行列化はそれぞれの一貫するタイプ2のエントリに関連する関連付けられたデータ内で識別され、内部リンクを通じて別のアウトバウンド・サブシステムへ送信できる。

#### マルチキャスト・ルーティング

以上、本発明のユニキャスト・ルーティングの態様について説明してきたが、第3図の転送用メモリおよび関連メモリ内のエントリと第7図の流れ図を再度参照しながら、マルチキャスト・パケットに関する本発明のルーティング機能について説明する。

本発明のMLDNEのマルチキャスト・ルーティングはMLDNE内でユニキャスト・ルーティングを実施する同様のハードウェア構造によってサポートできるが、マルチキャストはネットワーク要素の設計者に相当異なった問題を提示する。例えば、転送用メモリ内のタイプ2のエントリを導出するためのルーティング・プロトコルは当業者に周知のMOSPFおよびDVMRPなどのプロトコルを含む。これらのマルチキャスト・ルーティング・プロトコルはパケットのグループ宛先ネットワーク層のマルチキャスト・アドレスと送信元の送信元ネットワーク層のアドレスのループなしの分散ツリーを生成する。

MLDNEは、受信マルチキャスト・パケット・グループ宛先の第3層のアドレス、送信元の第3層のアドレス、およびインバウンド・サブシステムの到着ポートの機能としての、パケットを送信するためのいくつかの外部ポート（およびそれに対応するサブシステム）を生み出すマルチキャスト転送規則を備える。この依存性は、受信パケットのヘッダと突き合わせられる第3図の転送用メモリ内のタイプ2のエントリにそれぞれフィールド327、325、および337として反映される。代替経路で複製パケットを転送することを防止するため、到着フィールド337のインバウンド・ポートが含まれる。

受信パケットをマルチキャスト・ルーティングの候補として識別するため、MLDNEは少なくとも2つの判定基準に基づいてマルチキャスト・パケットを識別するように構成される。第1に、パケット・ヘッダは所与のクラスに一致する必要がある。第2に、パケットのヘッダはマルチキャスト・グループの宛先アド

レスを参照する既存のタイプ2のエントリに一致する必要がある。マルチキャストのケースでの一致するタイプ2のエントリはIGMPなどのマルチキャスト登録プロトコルの実行結果として作成できる。

第7図に、第4図のMLDNE201を介して受信マルチキャスト・パケットをルーティングする流れ図の例を示す。パケットがサブシステム410によって受信され、パケット・ヘッダが一定のクラスおよびマルチキャスト経路フィールド355を備えたタイプ2のエントリ321に一致し、ブロック703でこのエントリがマルチキャスト・ルーティング用であることが示される場合、制御は判定ブロック705へ移行する。パケットの生存時間を超過していない場合、受信

パケットのヘッダの生存時間フィールドをデクリメントすることでルーティング動作はブロック709でインバウンド・サブシステム410中で継続する。パケットのTTLを超過した場合、ブロック707でパケットはルーティングの代わりにVLANへフラッディングできる。一般にパケットのVLANはフラッディングに用いる第2層のトポロジ、言い換えると同報通信ドメインを定義する。

ブロック711で、インバウンド・サブシステム410は関連メモリ内のNEW\_VIDタグ・フィールド351に基づいて受信パケットの新しいVLANタグが必要かどうかを判定する。必要な場合、ブロック713で、パケットの第2層のヘッダのVIDが関連メモリにある次のホップの宛先VIDに置き換えられる。ブロック713は受信パケットの第3層のマルチキャスト宛先アドレスが同じVLAN内にある端末局を参照している場合に限って実行されることに注意されたい。このような判定はタイプ2のエントリの作成時にCPSによって実行された。

VLANがMLDNEによってサポートされているかどうかにかかわらず、ブロック715でインバウンド・サブシステム410は、第1の制御信号(sa\_replaceビット)をセットして転送元外部ポートに転送すべきパケットの第2層の送信元アドレスを外部ポートの送信元アドレスと置き換える必要があることを示して、パケットをMLDNEの外に転送する外部ポートにパケットをルーティングする必要性を通知する。ネットワークの層のヘッダ、特に生存時間(

TTL) フィールドを含む部分が変更されると、インバウンド・サブシステムはパケットのヘッダ・チェック・サムをブロック717で補正する。次にブロック719でインバウンド・サブシステム410は関連メモリのサブシステム・ポート・フィールド347で一致するタイプ2のエントリに対応するとして識別されたインバウンド・サブシステムの外部および内部ポートへパケットのコピーを渡す。

ブロック720で、パケットのコピーが内部リンクを通って異なったサブシステム420に到着した場合、動作はブロック721へ進み、ここでは分散フロー(DFまたはdistrib\_low)ビットと呼ばれる第2の制御信号がアウトバウンド・サブシステム420によって受信される。DFビットがセットされている場合、ブロック722で、クラス・フィルタはパケットのヘッダに基づいてパケット・クラスを決定し、タイプ2の検索(識別されたクラスを用いて)が実行される。

distrib\_low構成によってCPSはインバウンド・サブシステム内の一一致するマルチキャスト経路エントリに対応するアウトバウンド・サブシステム420内のタイプ2のエントリを定義することができる。この結果、CPSはさまざまな優先度をマルチキャスト経路をサポートする異なる外部ポートに割り当てて、MLDNEを通るパケットの待ち行列化の細分性をさらに制御できる。一致するタイプ2のエントリの関連付けられたデータ内のforce\_beビット(CPSが設定し、アウトインバウンド・サブシステム内のタイプ2の検索後に得られた)は、パケットとの内部リンク上で受信された優先度を無効にしてパケットが最低の優先度に設定され、外部ポートでの待ち行列化の細分性が確保される。

distrib\_lowビットがセットされていない場合、タイプ1の検索が転送用メモリ423で実行され、その結果、パケットは上記のタイプ2の待ち行列の細分性なしに転送されるかフラッディングされる。

一致するタイプ1またはタイプ2のエントリが見つかった場合、パケットは関連メモリ内で一致するエントリに対応するとして識別された外部ポートへ渡され

る。その後、動作はブロック 723 へ進む。したがって、インバウンドおよびアウトバウンド・サブシステムが異なる場合にマルチキャスト経路は C P S が生成する 2 つのタイプ 2 のエントリを必要とする。

ブロック 723 からブロック 729 の動作はサブシステム 410 またはサブシステム 420 であるアウトバウンド・サブシステムによって実行される。アウトバウンド・サブシステムは判定ブロック 723 で `s a_r e p l a c e` ビットがオンにされてパケットのそれぞれのコピーの第 2 層の送信元アドレスをパケットを MLDNE の外に転送するための対応する外部ポートの第 2 層のアドレスで置き換えるよう指示されているかどうか判定する。指示されていない場合、パケットは第 2 層の検索結果を用いて転送できる。

ルーティングのために第 2 層の送信元アドレスを置き換える指示がある場合、

ブロック 725 で、アウトバウンド・サブシステム、特にアウトバウンド・サブシステムの外部ポートはパケットの第 2 層の送信元アドレス外部ポートの第 2 層のアドレスで置き換える。次に動作はブロック 727 へ進む。ここで変更された第 2 層のヘッダについて CRC が再計算され、ブロック 729 でパケットが転送される。

以下に内部リンクを通じてパケットを送信し、情報を制御する革新的な構成および方法について第 8A 図および第 8B 図を参照しながら説明する。第 8A 図は使用されるパケット構成の簡素なブロック図である。より詳細に言えば、インバウンド・サブシステムがパケットに関するある情報、例えばルーティングを決定すると、この情報をそのままアウトバウンド・サブシステムへ搬送してヘッダ・フィールドの置き換えなどの後続処理がインバウンド・サブシステムが実行したのと同じステップを踏まずに容易に実行できるようにすることが有利である。さらに、終端間のエラーに対する堅牢性を維持することが望ましい。したがって、インバウンド・サブシステムはパケット 800 の周囲を制御情報 805 および巡回冗長符号 (CRC) 810 のカプセルに入れる。アウトバウンド・システムはカプセル化されたパケットを受信し、CRC 810 を用いてフレームの有効性を判定し、CRC 810 を取り除き制御情報 805 を除去してパケットを出力する

ための後続処理を決定する。

制御情報は必要に応じて出力前にヘッダ情報を更新する方法をアウトバウンド

- ・サブシステムに示すための情報を含む。本発明では、制御情報は以下を含む。

- ・ `replace_sa` — このビットがセットされると、ヘッダの送信元アドレス・フィールドをアウトバウンド・サブシステムの出力MACアドレスで置き換えるよう指示する。

- ・ `orig_tag` — これがセットされると、VLANタグがインバウンド・サブシステムに到着したパケットの元のタグであることを示す。

- ・ `mod_tag` — このビットがセットされると、インバウンド・サブシステムに到着したパケットのVLANタグが変更されていることを示す。

- ・ `don't_tag` — このビットがセットされると、`orig_tag` および `mod_tag` の状態にかかわらず、VLANタグを使用できないことを示す

(この実施形態では、通常パケットがCPS260から到着した場合に用いる)

。

- ・ `distributed_flow` — このビットがセットされると、パケットに関して初めて第3層または第2層の検索を実行する必要があることを示す。

- ・ `priority (2)` — 特定のパケットに対するサブシステム外部ポートでの待ち行列化優先度レベルを示す。

- ・ `reserved (9)`

第8B図は内部リンクを介して送受信されるパケットのヘッダ・フィールドの置き換えの処理を示す簡素なブロック図である。説明を分かりやすくするため、ヘッダ・フィールドの置き換えの実行処理に関係がないいくつかの要素は図示または説明していない。しかしながら、インバウンド・サブシステムはアウトバウンド・システムへの送信に先立って受信パケットを処理する要素を含み、アウトバウンド・システムは本明細書に記載する機能以外の機能を実行する要素を含むことは当業者には明白である。

第8A図を参すると、インバウンド・システム825はパケットを受信してデータベース（図示せず）を含むメモリにアクセスしてパケットに関する情報、例

えば、パケットをルーティングするか、またはVLANルーティングがサポートされているかといった情報を入手する。一定の制御情報が生成されて、カスケード出力処理 (COP) 835へ出力され、前記カスケード出力処理 (COP) 835は制御情報をパケットに付加し、付加した制御情報と共にパケットを出力インターフェース840へ出力し、前記出力インターフェース840はCRCを生成して付加し、パケットをカプセル化してアウトバウンド・サブシステム830へ出力する。好ましくは出力インターフェースはメディア・アクセス・コントローラ (MAC) である。ただし、他のインターフェースも使用できる。

アウトバウンド・サブシステム830は好ましくはMACである入力インターフェース845でカプセル化されたパケットを受信し、フレーム有効性チェックを実行しCRCを取り除く。入力インターフェース845はカスケード入力処理 (CIP) 850へCRCを取り除いたパケットを出力し、CIP850は制御情報を除去して、カプセル化に用いるCRCおよび制御情報がないパケットをパケット・メモリ855へ転送する。制御情報はメモリ855に記憶されたパケットに

対応する制御フィールド857に記憶される。出力ポート処理860はパケットを取り出し、パケット・メモリ855から制御情報を取り出して、制御情報に基づいてパケットを選択的に変更し、出力インターフェース865（すなわちMAC）へ制御信号を発行する。

パケットをルーティングする必要がある一実施形態では、OPP860がCRCに対応するパケットの最終4バイトを除去し、CRCを付加し、送信元アドレスを専用のMACアドレスで置き換えるためのMAC865への制御信号をアサートする。例えば、OPP860はMAC865へ送信するreplace\_SA信号を発行し、制御ワード内のno\_CRCビットをクリアする。別の実施形態では、VLANルーティングがサポートされる場合、制御信号の状態によってはOPP860はパケット内のVLANタグ・フィールドを除去し、CRCに対応するパケットの最終4ビットを取り除き、CRCを付加するための制御信号をMAC865へ発行する。より詳細に言えば、OPP860はorig\_tag、mod\_tag、don't\_tag、および第4の標識であるtag\_ena

bleを復号化する。tag\_enableはこの出力ポートに接続されたネットワーク・セグメントがVLANタグ付与をサポートしていないことを示す内部変数である。この変数は基礎のネットワーク・トポロジに基づいてネットワーク管理機構によって決定される。復号化処理の結果はOPP860がタグを取り除く必要があるかどうか、MAC865がCRCを生成する必要があるかどうかを示す。OPPは以下のテーブルに従って復号化を行う。

don't tag	tag_e nable	orig tag	mod_ tag	strip tag	regener ate CRC
1	x	0	x	Y	N
1	x	1	x	Y	N
0	0	0	x	Y	N
0	0	1	x	Y	Y
0	1	0	x	N	Y
0	1	1	0	N	N
0	1	1	1	N	Y

以上から、タグを取り除く必要がある場合、OPP860は好ましくはタグをMAC86へ転送する際にタグを取り除く。CRCを生成する必要がない場合、OPP860はCRCを生成しないことを示す信号（例えばset\_no\_CRC）を送信し、MAC865は受信したパケットをそのまま送信する。CRCを生成する必要がある場合、OPP860はパケットの最終4バイトを除去し、CRCを生成するための信号（例えばclear\_no\_CRC）がMAC865へ送信される。

MAC865は、OPP860から受信した制御信号に基づいて、送信元アドレス・フィールドを専用のMACアドレスで置き換え、パケットの出力時にパケットの最後に付加されるCRCを生成する。

カプセル化処理によって潜在的にパケットを数バイト拡張することができる。これによってリンクの要領に悪影響が出ることがある。この容量の損失を補い、標準のプロトコルが定義するよりも長いフレームの受信を可能にするため、プロトコル・パラメータ（本発明ではイーサネット・プロトコル）が微調整されて、プレアンブルのサイズが5バイト分縮小され、パケット間ギャップが5バイト分

縮小され、最大パケット・サイズが10バイト分拡大されている。

例示の目的で前述したMLDNE 201でのルーティング装置および方法の各実施形態は当業者の能力の範囲内でその構造および実施において変形形態をとることができるのは当然である。したがって、上記の説明は例示にすぎず、限定的なものと解釈すべきではない。

【図1】

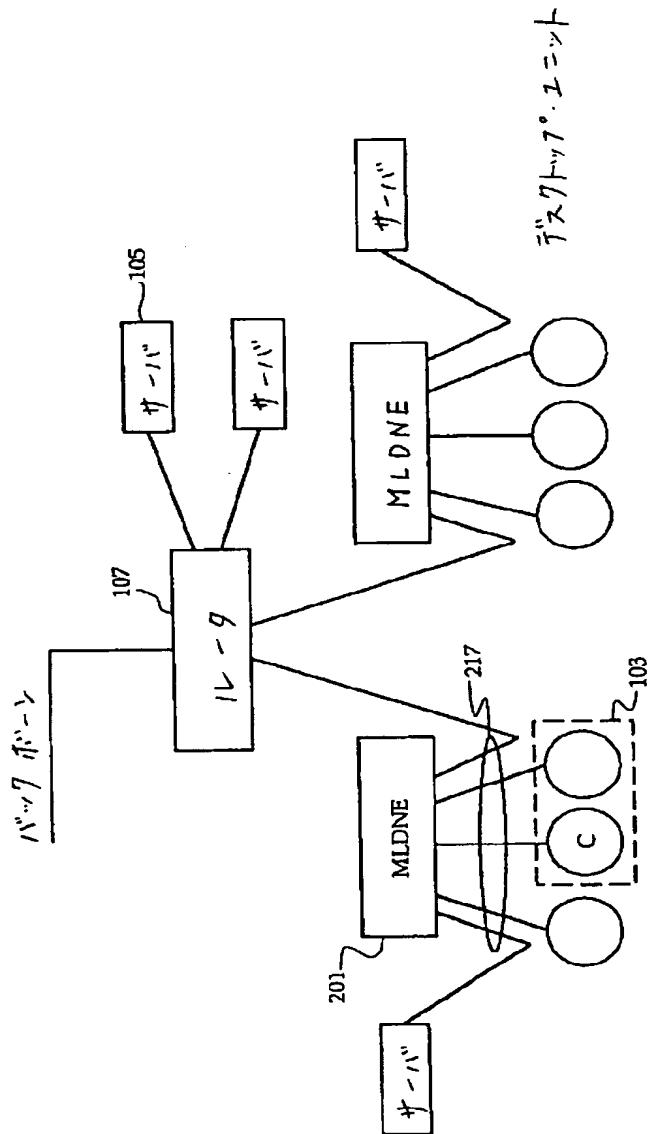


FIGURE 1

【図2】

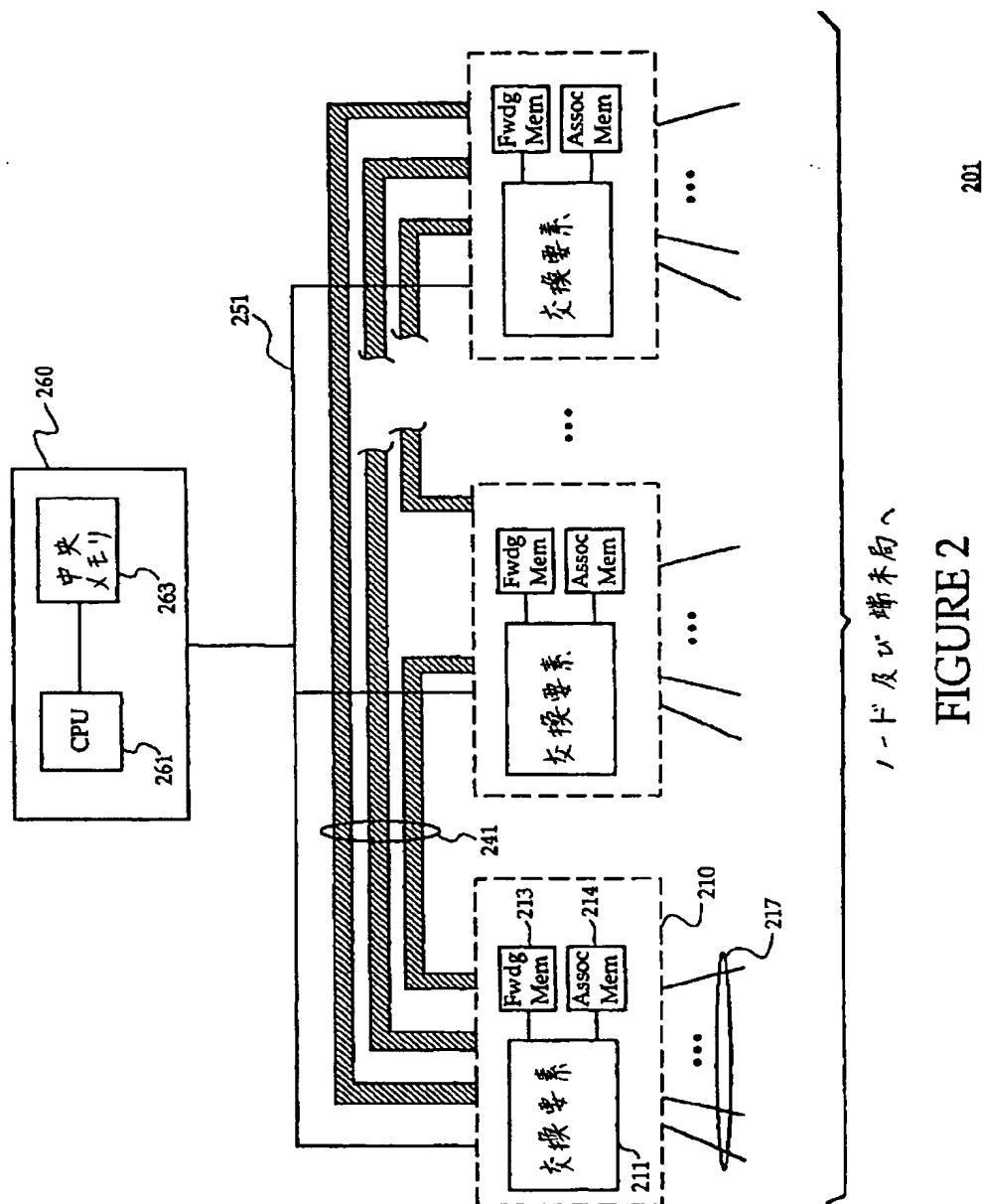


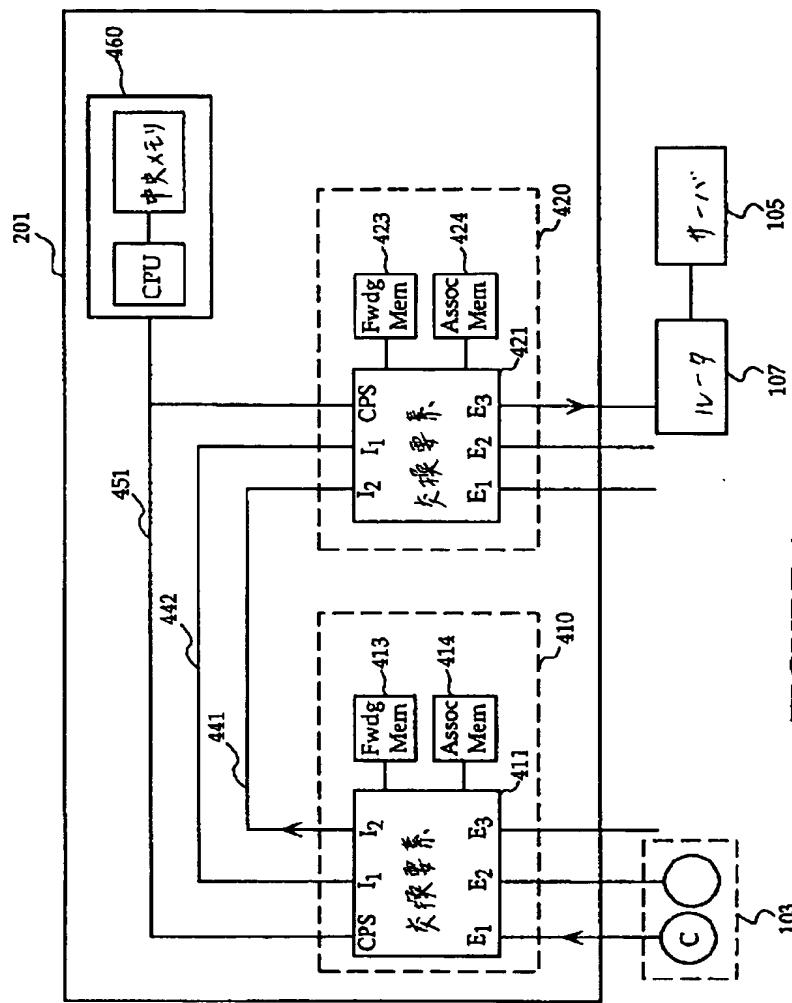
FIGURE 2

【図3】

転送メモリ (関連データ)		関連メモリ (関連データ)	
321		サブシステム データ	新ホード データ
322	タグ番号 送信元	第3層の 送信元	旧送信 元
323	タグ番号 送信元	送信元 データ	宛先 データ
324	タグ番号 送信元	データ	新ホード データ
325	327	333	335
326	327	333	337
301		サブシステム データ	新ホード データ
302	タグ番号 送信元	第2層の 送信元	VID
303	タグ番号 送信元	データ	VID
304	タグ番号 送信元	データ	VID
305	305	307	307
213		サブシステム データ	新ホード データ
214		サブシステム データ	新ホード データ

FIGURE 3

【图4】



## FIGURE 4

【図5】

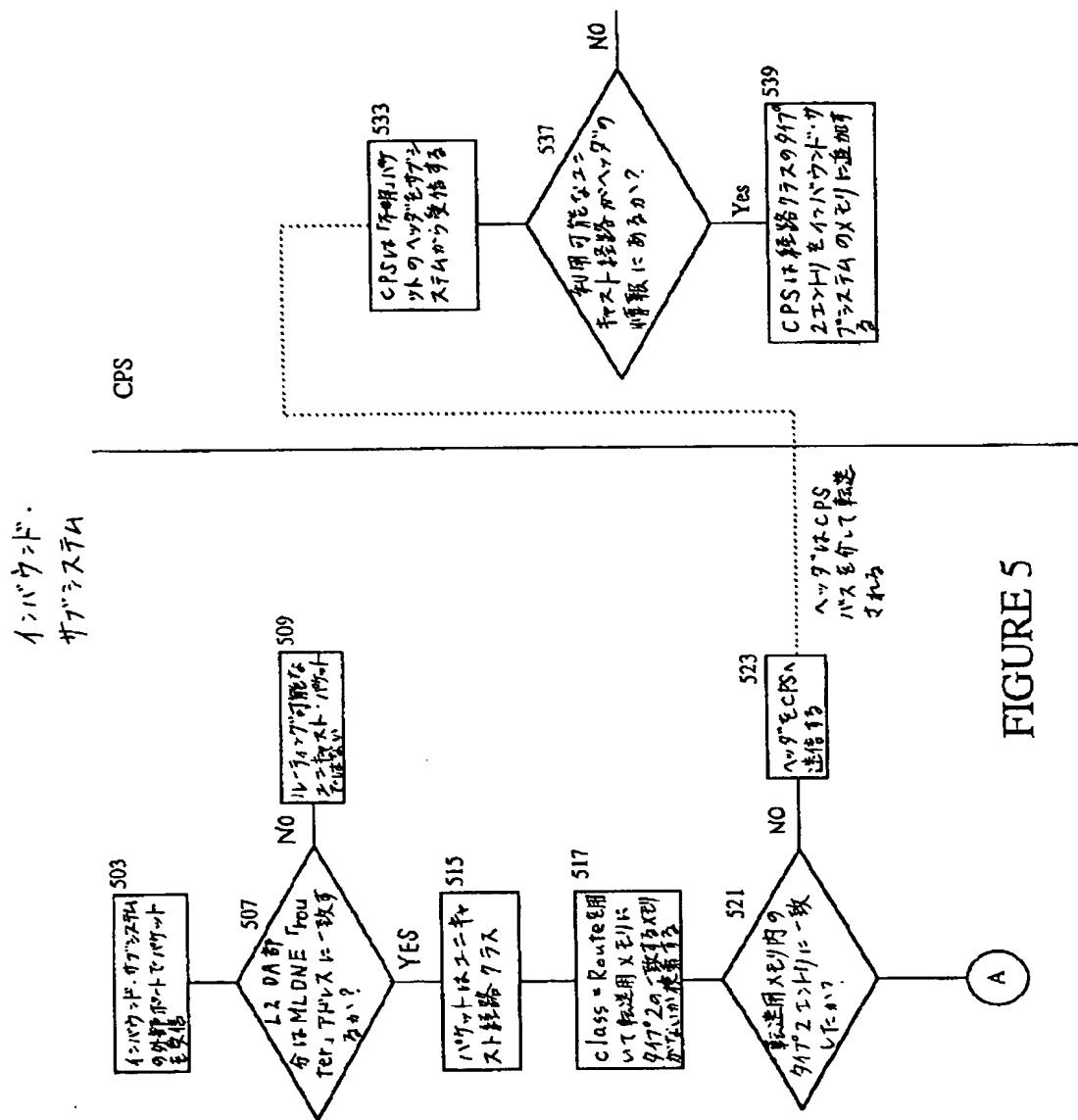


FIGURE 5

【図6】

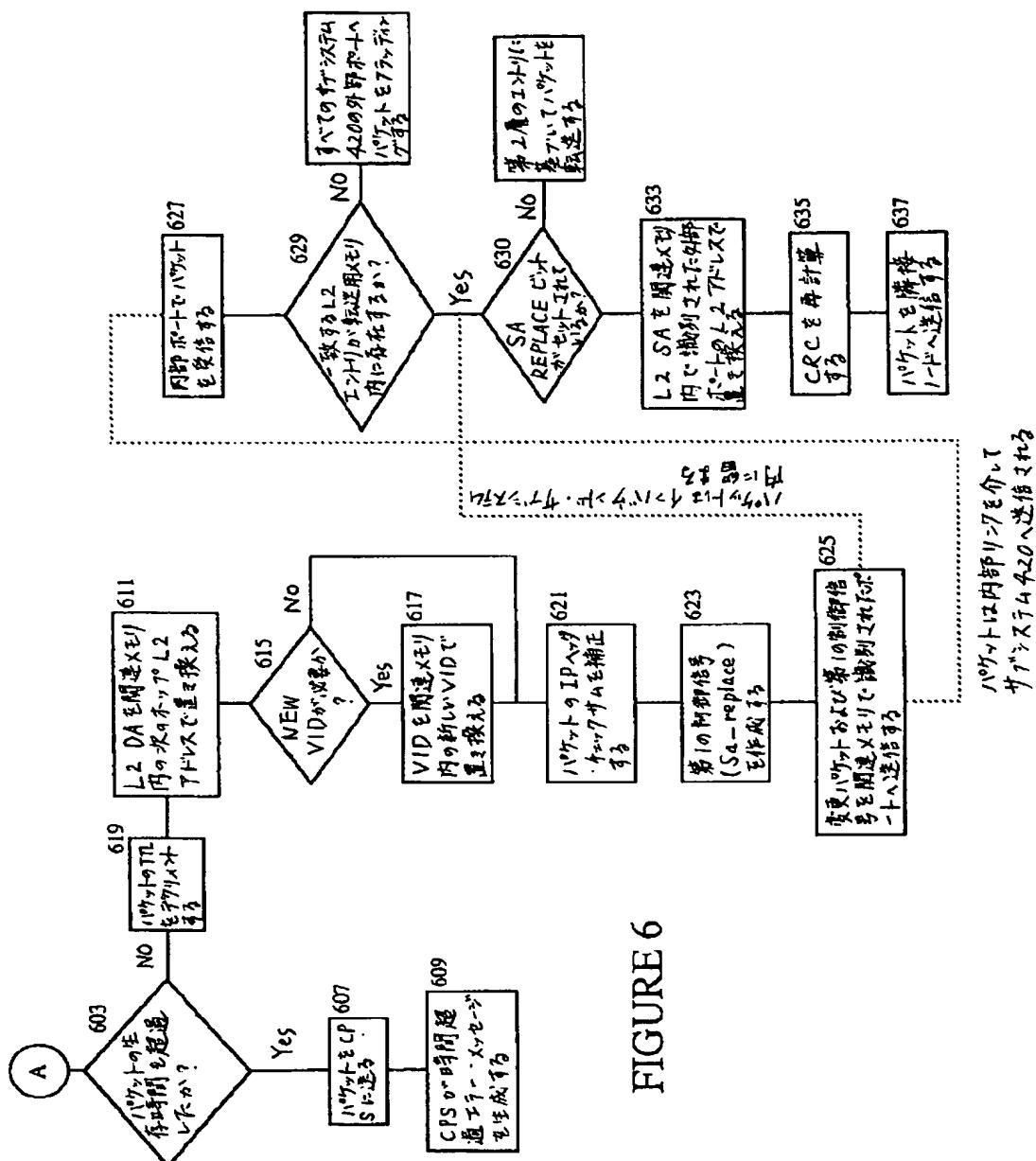


FIGURE 6

【图 7】

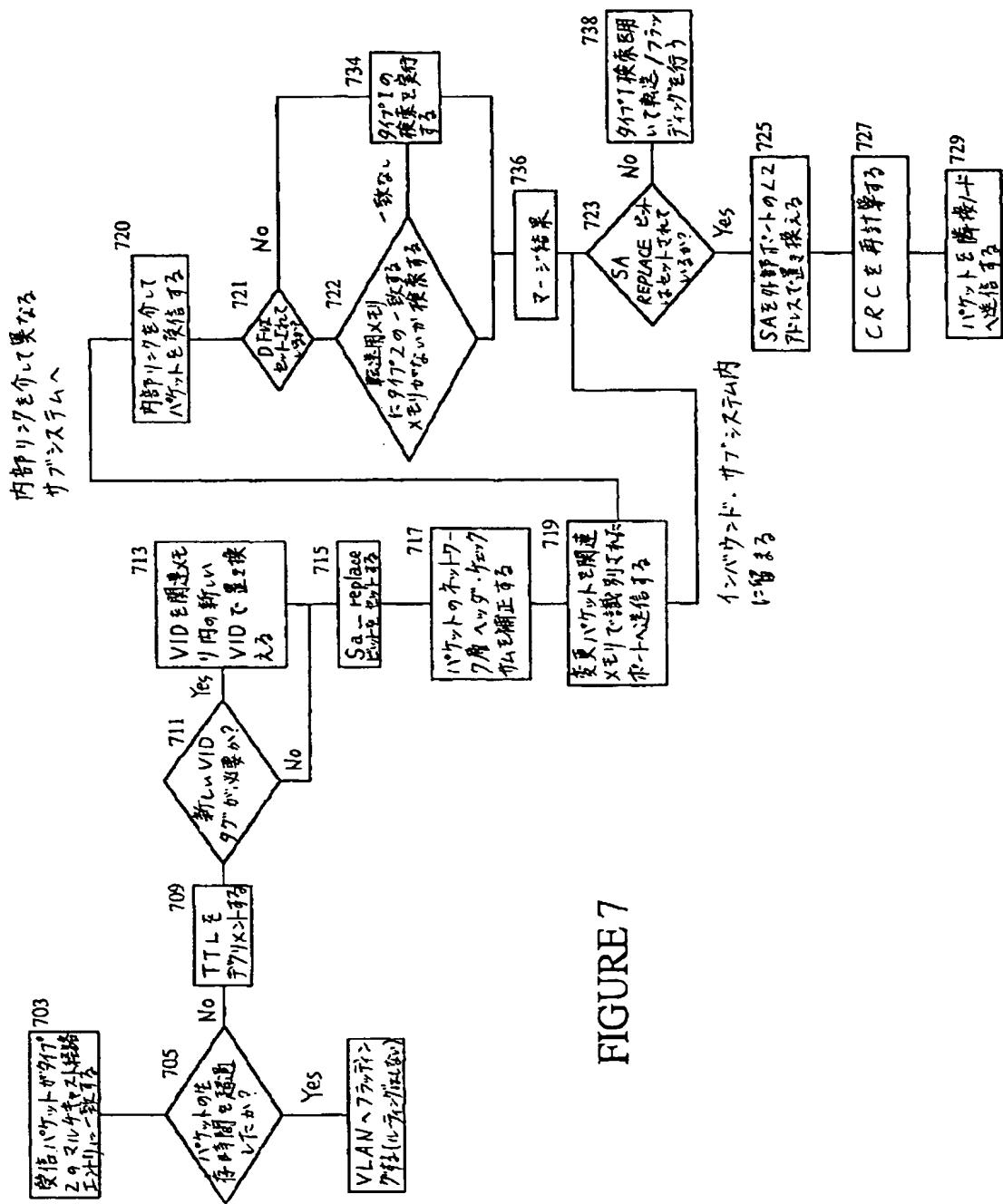


FIGURE 7

【図8】

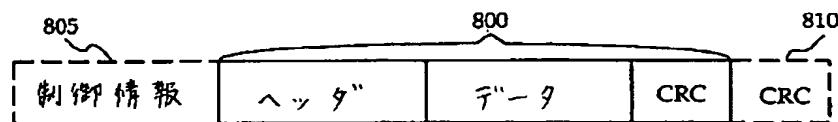


FIGURE 8A

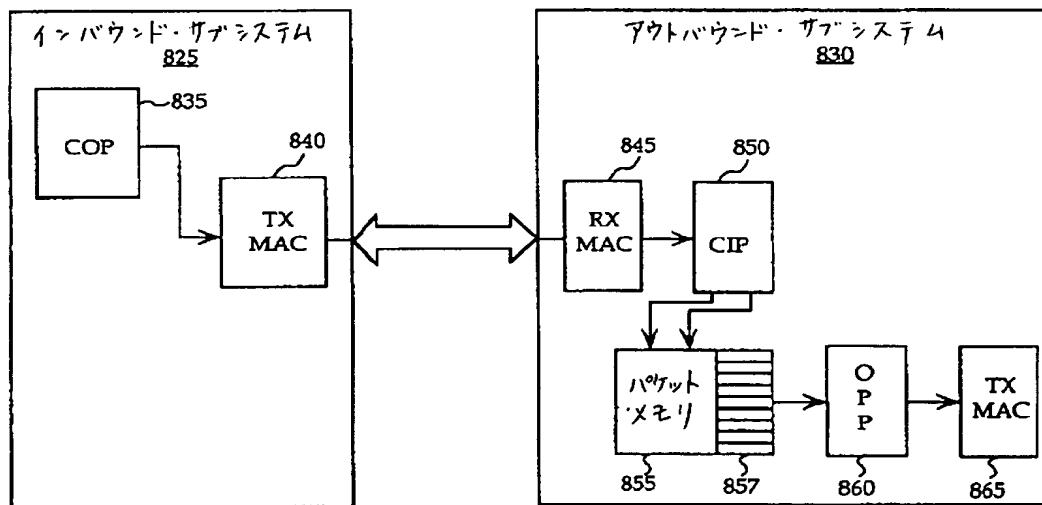


FIGURE 8B

【国際調査報告】

INTERNATIONAL SEARCH REPORT		International application No.. PCT/US98/13205																								
<b>A. CLASSIFICATION OF SUBJECT MATTER</b> IPC(6) : H04L 12/28, 12/56; H04Q 5/22 US CL : 370/390, 392, 400, 401, 402, 403, 404, 405; 340/825.52 According to International Patent Classification (IPC) or to both national classification and IPC																										
<b>B. FIELDS SEARCHED</b> Minimum documentation searched (classification system followed by classification symbols) U.S. : 370/390, 392, 400, 401, 402, 403, 404, 405; 340/825.52																										
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched																										
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)																										
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; padding: 2px;">Category*</th> <th style="text-align: left; padding: 2px;">Citation of document, with indication, where appropriate, of the relevant passages</th> <th style="text-align: left; padding: 2px;">Relevant to claim No.</th> </tr> </thead> <tbody> <tr> <td style="text-align: center; padding: 2px;">X</td> <td style="padding: 2px;">US 5,500,860, A (Perlman et al) 19 Mar 1996 (19.03.96), see col. 9, line 22 to col. 12, line 23.</td> <td style="text-align: center; padding: 2px;">1-18</td> </tr> <tr> <td style="text-align: center; padding: 2px;">---</td> <td style="padding: 2px;"></td> <td style="text-align: center; padding: 2px;">---</td> </tr> <tr> <td style="text-align: center; padding: 2px;">Y</td> <td style="padding: 2px;"></td> <td style="text-align: center; padding: 2px;">19-20</td> </tr> <tr> <td style="text-align: center; padding: 2px;">Y</td> <td style="padding: 2px;">US 5,592,476 A (Calamvokis et al.) 7 Jan 1997 (07.01.97), see col. 22, line 34 to col. 23, line 12.</td> <td style="text-align: center; padding: 2px;">19-20</td> </tr> </tbody> </table>			Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.	X	US 5,500,860, A (Perlman et al) 19 Mar 1996 (19.03.96), see col. 9, line 22 to col. 12, line 23.	1-18	---		---	Y		19-20	Y	US 5,592,476 A (Calamvokis et al.) 7 Jan 1997 (07.01.97), see col. 22, line 34 to col. 23, line 12.	19-20									
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.																								
X	US 5,500,860, A (Perlman et al) 19 Mar 1996 (19.03.96), see col. 9, line 22 to col. 12, line 23.	1-18																								
---		---																								
Y		19-20																								
Y	US 5,592,476 A (Calamvokis et al.) 7 Jan 1997 (07.01.97), see col. 22, line 34 to col. 23, line 12.	19-20																								
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.																										
<table border="0" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 10%; padding: 2px;">*</td> <td style="width: 40%; padding: 2px;">Special categories of cited documents:</td> <td style="width: 10%; padding: 2px;">"</td> <td style="width: 40%; padding: 2px;">later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</td> </tr> <tr> <td style="padding: 2px;">"A"</td> <td style="padding: 2px;">document defining the general state of the art which is not considered to be of particular relevance</td> <td style="padding: 2px;">"X"</td> <td style="padding: 2px;">document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</td> </tr> <tr> <td style="padding: 2px;">"E"</td> <td style="padding: 2px;">earlier document published on or after the international filing date</td> <td style="padding: 2px;">"Y"</td> <td style="padding: 2px;">document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is taken alone</td> </tr> <tr> <td style="padding: 2px;">"L"</td> <td style="padding: 2px;">document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</td> <td style="padding: 2px;">"&amp;"</td> <td style="padding: 2px;">document member of the same patent family</td> </tr> <tr> <td style="padding: 2px;">"O"</td> <td style="padding: 2px;">document referring to an oral disclosure, use, exhibition or other means</td> <td colspan="2" style="padding: 2px;"></td> </tr> <tr> <td style="padding: 2px;">"P"</td> <td style="padding: 2px;">document published prior to the international filing date but later than the priority date claimed</td> <td colspan="2" style="padding: 2px;"></td> </tr> </table>			*	Special categories of cited documents:	"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention	"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone	"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is taken alone	"L"	document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family	"O"	document referring to an oral disclosure, use, exhibition or other means			"P"	document published prior to the international filing date but later than the priority date claimed		
*	Special categories of cited documents:	"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention																							
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone																							
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is taken alone																							
"L"	document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family																							
"O"	document referring to an oral disclosure, use, exhibition or other means																									
"P"	document published prior to the international filing date but later than the priority date claimed																									
Date of the actual completion of the international search  14 SEPTEMBER 1998		Date of mailing of the international search report  19 OCT 1998																								
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230		Authorized officer DUONG, FRANK  Telephone No. (703) 305-5428																								

---

フロントページの続き

(72)発明者 ミュラー, シモン  
アメリカ合衆国・94086・カリフォルニア  
州・カッパチーノ・ラ メサ テイラー  
ル・983

(72)発明者 ザウメン, ウィリアム・ティ  
アメリカ合衆国・94303・カリフォルニア  
州・パロ アルト・クララ ドライブ・  
912

(72)発明者 ヤン, ルイーズ  
アメリカ合衆国・94070・カリフォルニア  
州・サン カルロス・ロジャーズ アヴェ  
ニュー・110